Dennis Jennings: The IDN Variant issues project – what we are now beginning to call the IDN VIP – variant issues project. And we're herethis afternoon to launch the project and to tell you a little bit about it and how we intend to proceed. Kurt Pritz was hoping to be here, he's the executive responsible for the project, but he is triple booked at the moment and he's the sole speaker in another session so he's not here. So I will go straight into the first presentation and ask Francisco Arias here to put up the first slide which gives us the agenda, which you probably should have. You see that we have a number of speakers.

I would just like to introduce the team – the project team – Kurt Pritz is the executive sponsor, I am the project leader, and the team members are Anand Mishra, Naela Sarras, Kim Davies, Baher Esmat, Steve Sheng and Francisco Arias who are all here. So that's Francisco, that's Naela, that's Kim, we've got Steve here, Steve, Baher. Yeah, we're all here. Good.

And we have a number of presentations – could we look at the list of presenters who will all be speaking for presenters. But first, I'm going to ask Francisco to outline the project proposal to you. It's a draft. It's available for public comment and we're looking for feedback on the project proposal so that we can finalize the proposal and kickoff the project. So Francisco, over to you.

Francisco Arias: Hello. So, I was talking about why we are doing this break. As some of you may know, this is a long standing request from certain members of the IDN user communities for several years. we are stuck in this break now because there is a Board mandate from the resolution on the 25th of September 2010 to start a project to basically define what the problem is to develop an issue report.

The current status today is that the applicants made it clear the variant strings for the TLD that in their application, this is for the new gTLDs and in the IDN Fast Track, but no variant TLD string will be related until the proper solutions are developed and implemented and this will be the first step in that direction.

So, what has been proposed in this project is to conduct five study cases with participants from the different communities. In this case we're proposing Arabic, Chinese, Cyrillic, Indic, and Latin. And each of these study cases will be composed of community members experience in DNS, security, policy, linguistics, registry operations, and community representatives. We are thinking that maybe some of these members may be shared between the teams given that some of these specialties, let's say, are of global nature and do not need to be from the community. For example, DNS is the same no matter where.

We also are aware that this is not the first time that this has been done – there are previous works on the IDN variants and we intend to use that as an input. We also know that there are coordinating works working on this subject. For example, the joint working groups on IDNs from the ccNSO and GNSO is working on policy for IDNs in general and one of the topics is the variant TLDs. We have a session, we were in their session on Monday and we talked about the subject and see how we could coordinate our reports. There is also technical forming their way in the idea in the DNSEXT Work Group, trying to define what is called their name aliases, which is something that could be used eventually as a solution for the variants.

In terms of the project scope, as I mentioned before, the idea is to define the problem and we are not starting a break to actually produce the solutions, we are trying to define the problem. Why we are doing this – when people talk about variants normally, there are different things that people mention. So one of the first things we need to do is clearly define the terms and get with the technical and linguistic communities so we know what we are talking about. And another goal of this project will be to identify the challenges, you could say the requirements for the IDN variant TLDs.

The final outcome of this project is expected to be an issues report integrated the common issues and the case specific issues for each of the study cases. We also are thinking of having intermediary reports

from each of the case studies. This is the proposed timeline. Important item here is the recruitment phase where we would like to have participation from the community. We're expecting to have it done by the end of April.

And as I mentioned, the case studies would be something in the middle of the project, by the end of September and we are expecting to finalize this project by the end of this year. We understand that this is something that some communities really would like to have solved by yesterday, we will try to work as fast as possible, but given the complexness of the problem this is the timeline that we think is doable.

Finally, we would very much appreciate your feedback. The project proposal is currently in public comment. There you can see the link. The public comment ends on the 6th of April. And if you also, you could talk with any of the project team members while we here in San Francisco.

Dennis Jennings: Thank you Francisco. Can you bring up the slide that shows our definition of variance because I think that's an important – now I don't want to get into a long discussion about variance, but I want to be clear what we are taking as our definition of variance as our starting definition. So, we're talking about variance we're not talking about confusability. We're talking about "variant characters occur when a single conceptual character" – some

people suggest it should be named slightly differently – "but a single conceptual character can be indentified with two or more different Unicode Code Points with graphic representations that may or may not be visually similar. IDN variant TLDs contain one or more characters that have such variants."

That's our definition and that's what we're working with. I know that there are a variety of definitions, but that's the one we're starting with. So, thank you for that. Could you put up the slide in relation to the Cyrillic case Francisco? Thank you.

Now, as you saw from Francisco's presentation we had identified five study cases – Cyrillic, Arabic, Latin, Chinese and Indic. Thank you. We've already been advised that Indic is such a broad term with such a vast array of scripts that we may have to focus a little more closely to actually get any reasonable work done.

But in the case of the Cyrillic script, we consulted with Dr. Andrei Kolesnikov, who is the CEO of the Coordination Center for .ru, and he has advised that IDN variants are "not applicable in our case". So, we have that advice. Now that doesn't mean that that is necessarily a true statement, but that's the advice we've received. We will publish that on the website and we look forward to comment on that and advice on that from the community.

If that is a true statement – there are no variants – if you look at the script table it looks obvious, then maybe this part of the project is quite brief and we might move on to the confusability issues in relation to Cyrillic, Latin, and Greek. But make no immediate jumps to conclusion and just simply say that's the advice we've received and we put that up on the website and see what the community says. So that, for the moment, deals with the presentation on the Cyrillic case.

And id' like to ask our first presenter, Dr. Sarmad Hussain, who's Professor and Head of the Center for Language Engineering, the Institute of Computer Science in Lahore Pakistan. Francisco has your presentation. We'll sort that out in a second. We need to get a new presentation because there are animations in the presentation. Sorry. So Dr. Hussain is going to talk to us about the Arabic IDN variant case study. Thank you.

Dr. Sarmad Hussain: Thank you. So the aim of this – I've been given 10 minutes – so it's a reasonably wide topic and I'm going to go through this reasonable quickly and I'm not sure if it's probably not possible to explain the whole thing in 10 minutes. The idea is to give you just a sense of what kind of issues exist and obviously if you have any questions, more detailed questions, feel free to come and ask after the session.

So, Arabic script is used across a reasonably wide geographical region. It's used to write more than about 80 different languages. And here are some examples of the languages which are written in Arabic from the different regions. Basically Arabic writing system is a consonantal writing system – what that means is that only the consonants are written; vowels are actually not explicitly written unless they are long vowels. And short vowels are optional, which means the user or a writer may choose to write these or may choose to not write these or ignore them.

Another very interesting thing about Arabic script is that is bidirectional, which means that letters are written from right to left, but digits are written from left to write. And this will show you an example of each letter appears towards the left of the previous one, but the digits appear right of the previous one. So when you are writing digits and letters the direction of the writing system changes.

Another thing, which you probably saw, was that it's a cursive writing system which means that the shapes of the letters join with each other and shapes of these letters actually change based on whether they're initial part of the, or medial or final part of the joined form, which is normally referred to as [allegation]. And then there are certain characters which don't join at all.

Alright, and they are represented in Unicode in two areas 0600 to 06FF and then 0750 to o77F. And they're a lot of different writing styles which are used – mostly they're two traditions – the Naskh style, which is the first one, and then Nastalique style, which is the second one. The others are mostly stylistic, but the first two are actually used very completely by different language communities in the world.

Okay, so given that background – the reason I gave that background to you was because I will probably refer to some of this later in my presentation. There are two kinds or sources of variants which are caused within Arabic script. One which I'm calling intrinsic is because the community which uses the language and the script consider two different strings to be the same even though they may visibly be different.

The other reason a variant is caused is because something which is coming from within the language or script community, but because the way the script has been coded by the Unicode standard. And that's what I'm referring to as an extrinsically motivated variant. The extrinsic ones are mostly - look exactly the same. The intrinsic ones may or may not look exactly the same.

So, why do we want to have variant management? There are two very significant reasons. We need to protect users from security threats, like

phishing, and we need to give perception to the users that the internet, whichever the way a type of string still resolves to the same URL and that is only possible through a variant management process.

And I just want to very briefly touch on some of the linguistic issues and again, this particularly more detail just come back and talk to me. So there are technical issues, there are user interface issues, and there are policy kind of issues which need to be addressed. Just to give you a examples – one of the exact variant issues is caused through the normalization process that the way Arabic is written there is a code character and there is a combining mark sometimes and sometimes what happens in Unicode is that you get both versions.

So you get – there is a Unicode for the combined form and then there is separate Unicode's for the uncombined forms which can actually combine to give you the same string. And some of these normalized forms are defined by Unicode, but actually some of them are not. So I will just give you an example of these two characters at the bottom. If you look at the second row, one of these characters is 691 and the other one is a combination of 631 and 615. Can you tell them apart? If you cannot, then if they're not mapped onto each other then somebody will be able to phish this site which will be using this character.

Another issue occurs, a similar issue occurs because vowel marks, as I said, are optionally written. So, that's the second line at the bottom – you will see if I combine 627 with 64F, the 64F is an optional mark, even if I don't write it it's considered the same as when I write 627 without 64F. So, the – let me actually do it this way, it's not showing it properly. So the first two strings are the same – the one with the vowel mark and without the vowel mark.

The second two strings are the same as well – the one with the vowel mark and without the vowel mark. But interestingly, if you put the vowel marks, the first and the second strings are actually not the same. So, the first and second map out to the same string as far as [equivalence] is concerned, but if you make it explicit they don't map onto each other. So that actually causes a problem.

So there are then this third kind of similarity where you have actually some different characters which sometimes actually look like the same thing. So as I said, Arabic script has four possible shapes of a character. A particular shape may be distinct from another particular shape of another character, but in another context those two shapes actually may look very similar. So if you look at the medial form of these two characters, the isolated forms are totally different. So there are two distinct characters, but if you look at the medial forms or the initial forms, they are exactly the same. I've given you an example of an IDN ccTLD string .pakistan – and one of the strings is written with 06A9 and

the other one is written with 06F3 and can you tell the difference? You probably cannot because actually there is no difference. They're identical to each other even though they have different codes behind each other. So that's a variant case.

I give you an example of an exact match – there are also some approximate matches, so there are other characters which are available in the IDN, in the Arabic table, which two users – they may look different, but users may perceive them as the same characters, only stylistic variations. So something like a difference between a Time New Roman font and Arial font – users may perceive a similar difference between 6A9 and 6A8, the two different [cuffs].

So these six different string which I've listed here, for .pakistan for example, may actually be perceived as the same string for the people who use the string in Pakistan, even though they have come from a variety of, at the back they have a variety of unique other code points.

Moving on – there is a real issue with digits. Obviously people use ASCII digits now freely, mixed with Arabic digits and Arabic itself has two sets of digits and Unicode will tell you the reason why. And so the question then is – I've listed some strings here – are all these strings – so all these strings which I've listed are composed of different Unicode's. So at the back, at the DNS layer, they are different strings technically. Do you

agree that they're all different string or do you think they are the same – or some are different or some are same? So who decides what's different or same? And that's one of the issues which obviously this project needs to resolve.

Then there are also some other issues. So I'm going to skip some of these things, you can look at the presentation or you're free to contact me later. So there's some of these technical issues. There's also some application layer issues. So I've written the same, exactly the same strings in Internet Explorer and Google Chrome and you can see that the sequence in which they're written is one is left/right and one's right to left. And therefore it is not very easily possible for a user to understand what's actually written. So there's inconsistent user interface. Also the user bar is very small so if you use a mark it is very hard to tell which mark it is because it is such a small font. So it can also cause phishing kind of issues.

In any case, those are some of the technical or application layer issues. There are also obviously, once we start talking about variants, there are some policy level issues, meaning how do we actually articulate these variants in a language table. There are no formal mechanisms at this time which are defined. How do we implement these variants in registries; whether we bundle them, block them, reserve them. There are again, not enough works in this area which is, but there is work being done in these areas by different groups. So obviously we would

want one world – one internet; that's the slogan we have. But if we do want one internet, variants have to be handled. Otherwise we will not actually have one internet. So it's a very fundamental issue which needs to be addressed. Thank you very much.

Dennis Jennings: Thank you very much indeed for managing, in 10 minutes, to cover a huge topic for us. So that gives you some indication of the level of complexity in just, in the various scripts. And our goal is to try and define what the user expectations are, what the requirements are for the solutions. So thank you very much indeed. Could I ask our second speaker, Dr. Xiaodong Lee, who is the VP and CTO of CNNIC, to come up and tell us about some of the user requirement issues for the Chinese IDN variants. Sorry, thank you Naela. I'm going to take questions at the end. I'd like to get through the presentations and then take questions at the end. Thank you.

Dr. Xiadong Lee: I'll use Francisco's computer to make my presentation. Upon Dennis's request this is a 10 minute presentation – I cannot give you much information about what's the variant issue by the Chinese domain so I just give you very brief introduction about what is the Chinese variant and what is Chinese variant (inaudible) and what's the policy issues. Next slide.

It's very small, my fault. You know, Chinese languages throughout the world are more connected to China and not only China, to Chinese culture than before. I will give you some statistics of two years ago, but it was published by the government. There are over 100 million passengers go from Taiwan, Hong Kong and Macau they also Chinese speaking area to mainland China, and their (inaudible) study meaning from mainland China to Hong Kong, Macau and (inaudible) – that's the telecommunications with each other. Also there are so many Chinese around the world except for the 1.4 billion Chinese users in Mainland China, Taiwan, Hong Kong and Macau, there are about 48 million Chinese users living in other countries, including the American and the Europe. Especially you see so many Chinese people in San Francisco.

For the Chinese variant issues that the Chinese have two writing forms – one is simplified Chinese and the other in the traditional Chinese. Simplified Chinese is primarily used in Mainland China and in Singapore – I think it's the official language form. and of course you can see some Chinese, simplified Chinese characters in Japan. For the traditional Chinese used primarily in Taiwan, Hong Kong, and Macau and the Southeast Asia countries; and also used by the Chinese communities origin in other countries.

Now the SC and TC are recognized I think as interchangeable and millions of Chinese users use both TC and SC in their daily lives

and if any attempt to separate the SC and TC could create user confusion and maybe result in some critical issues. But if we enable this as it's intended to be we offer the tremendous convenience for the users. And I give you an example in San Francisco that is was said that over 200 thousand Chinese people living in Bay Area – from the news I heard more than 200 thousand peoples. Most of them are speaking Cantonese and writing traditional Chinese because there are so many immigrating from (inaudible), from Hong Kong and from other areas speaking Conges. But in recent 30 years, more and more people speaking Mandarin and writing in simplified Chinese because there are so many immigrating from mainland China that it means that San Francisco in the Bay Area is mixed TC and SC.

Simplified Chinese characters correspond to more complex traditional characters. What is the simplified Chinese and what is the traditional Chinese – it's a little bit difficult to define, but it could refer to the RFCs. IFC 3743 is made by [GET] to join engineering team; this team includes experts from China and from Japan and from Korea. But specifically to Chinese, for Chinese to (inaudible) we have IFC 4713, only for Chinese (inaudible) and institutions. Unicode, latest Unicode standards, there are over 70,000CJK Unified Chinese characters. At last look I only can recognize 10% of that. So it's a very big user to open that kind of, so many Chinese characters for Chinese domain registrations so

we define only about 19,000 characters open for Chinese domain registration.

You can reference to this link. It's published by (inaudible) for simplified Chinese and traditional Chinese. In the 19,000 Chinese characters less than half is less than each (inaudible) Chinese character have one or more variants, you can see that by the issues. Another information for you is that the Chinese variant issue doesn't mean only TC and SC variant – we also have other kind of variant issues except for the simplified and traditional Chinese.

Okay, just a quick example, you can see the two Chinese domain names. That means the Bank of China.China. It's a real name; it's not only an example, it's the Bank of China is the most famous bank in China. But if there is two different registrars here there would be some problem. Okay, next slide. If only Bank of China in simplified Chinese only allowed to be registered – so that will bringing much trouble for users. You can see the picture in the left, down there, it's the logo of Bank of China.

You can see that now is the current logo, it's the current logo of Bank of China. It's the logo of Bank of China now totally is a traditional Chinese character. That means that even now, the even in China so many companies organized users are traditional Chinese as their brand name logo. And the left side is the Chinese Academy of Science – it's my boss, my organization. There is also

used the mixed Chinese characters. The second one is the traditional. And the fourth one is the simplified. The traditional one is not this one. Okay, just give you an example. Next slide.

So what's the requirement for Chinese IDN variant TLD? That means if applicant applies, a Chinese applicant apply for a domain the language tag should be provide. If it is Chinese, both simplified Chinese only and traditional Chinese only forms must be added to the root servers and mixed variants should be reserved for any phishing or other security issues considerations. But what's the issue for CJK? Next slide.

In China, Japan and Korea they use the Han character, that means Han character. You know, we still separate the language into the Chinese and non Chinese – the Chinese is ideographic language in the world now, but others they think it should be alphabetic or phonetic. So I think that Chinese only ideographic script is being used by native speaking population today.

So we have worked for Chinese issues for over 10 years, we have published two RFCs for the Chinese name registrations. So also in Japan they use Conge is Han character and in Japan they use Hanja for Chinese character, but in Japan they only use simplified and in Korea they only use the traditional one. Next slide.

So that means that Japan and Korea, they don't need variants. The variant issue is only an issue for the Chinese community. So I also refer to two links – they were published by the [GPI] and (inaudilbe) .kr and .jp. You can see, check the table, they are only variants in this table. Next slide.

Okay, it's the last slide for my presentation. I think that the common issues for the variant issue is application issues and other issues, but I want to mention some special issues for Chinese. The first potential issue is for the string similarity evaluation. Maybe for the people they only need 3000 Chinese characters, but there may be 6000 that are popular. So many people can recognize 6000 but only use 3000. But now, for CNNIC members, CNNIC and [inaudible]NIC and other CNNIC members, they open over 19000 Chinese characters for Chinese domain registration. But you know that also have other more than 50000 Chinese characters is the potential registration.

So what's the limitation for registration? How many is enough? That should be asked. And I also give you two examples that is different forms but similar meaning – two characters that means "eat". But you can see that the two pictures are totally different, but the meaning is the same.

Another employs different meanings but is similar fonts. You can see that. This first one is "sun", it means "sun" – the California

sunshine right? And the second one means "say" – you can see that the picture is very similar. So the (inaudible), similarity variation is very important. Another issue is the delegation combination. Now for .china and .[havan] they have two pairs, for two users now. But it was for two pair, two from the end user and the second level is top level, it would be combination issues. But yet the principle is better than nothing.

Now with paired delegation we can give some improvement about the registration and administration about the Chinese domain name, but it is not perfect, but it is better than nothing. But we need a long term solution. So understand that here. It is under discussion, a long terms solution for the many variant issues management. But maybe we shouldn't wait 10 years for adopted by the whole world. That's my consideration.

Dennis Jennings: Thank you very much indeed. Thank you very much for that presentation. Again, we move on and we'll take questions at the end of the presentation, but we're beginning to get a sense of the complexity of the problems that we're trying to work with. Our next speaker is Dr. Mahesh Kulkarni, who is Program Coordinator and Head of GIST Center for Development of Advanced Computing in India. Dr. Kulkarni.

Dr. Mahesh Kulkarni: Thank you very much. As you can see the screen basically so many scripts are written down and the message is "welcome",

written in all different languages. The keypad label showing different languages being used in India. This is current, what is called the facts and figures and multilingual diversity of India. Constitutionally 22 languages are there in India. There are a total of 45 living languages out of which 22 are constitutionally recognized languages.

Two major script families are being used – one is Perso-Arabic family and also Brahmi based family, which means that some of the languages get written from right to left and some of the languages get written from right to left and left to right. Especially Sindhi, Kashmiri, Urdu uses the Perso-Arabic system with notational changes in Sindhi. The remaining 19 languages use 11 derivations of the Brahmi script. So there is one to many and many to one relationships between the language and script. For example, Santali and Sindhi – these are two languages – they use more than one script. Devanagari, which is also a writing style, is used for Sanskrit, Hindi, Marathi, Nepali, Konkani, Maithili, Dogri and Bodo language. So this is a relationship where one to many and many to one relationship exists.

These are some of the current issues – I can say the challenges I will point out. We have got several alternate spellings that the same different spelling is been written differently. Phonetically they are the same. So whether we are to treat them as a variant or we don't want to treat them as a variant – that is one part.

Secondly we have alternate forms also. So the same spelling is being written in two different ways – Hindi and Hindi – the both ways that you can see. And then there are issues of something called a different inputting mechanism as reordering levels.

This is what exactly what is called the fonts look like. So you order some base character and on top of which some what is called the (inaudilbe), and you see the complexity of the Indian language all here. And you also see the Perso-Arabic family which is getting written from right to left and we would (inaudible script of writing. And when the Conges come in the picture, the resulting glyph shapes increase manifolds, which means that if there are two characters which are getting combined, the third resulting character will have a totally different shape altogether.

Then we tried to define what is something called the types of variants basically, so we sought some definition of the variants over here. Now, one of the definitions what we feel like we'd like to treat is homographic variants that is similar looking. The example of Latin 1 and l at the smaller point size and the address body looks similar. Same things happen in the Indian language – like the one is a (inaudible). They look almost similar in smaller font size.

The second variation is something we are calling eh homophonic variants that is similar sounding or alternate spelling. So as these

are the examples or color and colour in Latin, the same thing applies over here Hindi and (inaudible), and both are what are called the permissible forms and alternate writing styles. The case variant especially is not the case is not over here because there is no capitalization in Indian languages unlike Latin or Roman.

Confusingly similar – so what basically is happening is that in most of the browsers and applications use Idn display labels in minimal size. This results in a maximum number of spoofing and phishing attacks. Muti-tier scripts such as used in Indian languages are less readable in the address bar. And Unicode normalization rules have also been considered as variants. So which necessarily means that two different Unicode Code Points can what is called have one shape as well as you can have a different Unicode Code Points having the similar shape coming up.

This is what exactly we are calling a homographic variant – that is a similar looking at the smaller font size. If you really see this is the Telugu script and in the first box and the second box, if you see, just the dot which is blue – the character, otherwise they look similar. So this we will treat that something called a homographic variant because at a smaller point side that dot will not be visible at all. In the case of Tamil variants at the bottom, you can see that when combined by two different characters, while you can see that we have one Unicode Code Point which is 0B94 and they will exactly the similar.

Then homophonic variant and alternate spelling, which I discussed about, homophones are there – Hindi versus Hindi – so this is a different way of writing but both mean the same. Like Hindi with bindi and (inaudible) also looks same. Common misspellings are also one of the major issues over here - .India and India – and while formulating the IDN policy for .in we have not considered this variants as historically other domains have always considered alternate spellings of the color colou.com as separate entities. So we followed an approach wherein we are not going to treat homophone as a variant, however we will be treating homograph because that is especially what is getting displayed into the user browser.

Case variants are not applicable in case of Indian languages. However Indian languages are rich in synonyms. So when you take a word like "parrot" it has got a synonym like Indian, Hindustan, and so many synonyms are there. And we look forward for some solution towards that.

There is a major issue in terms of some of the invisible characters like the zero with joiner and without joiner when used within the (inaudible), it might be reassured here that again same, the end user may not be able to decipher but the rendering engine in the same way. Need for a variant identification – Indian scripts introduce syllabic variants. So that is something called the syllabic

variants. As you can see but there are two examples at the bottom which I showed to you and these are called the syllabic variants and they actually mean the same.

This is a classic example where I show three cases, [Mahastara] written in three different if you seen what is called there the bottom right – with zero with joiner and without zero with joinder and transferable but all of them look similar. IDN variant TLDs – so what we were suggesting is that we cannot just translate to .com because .com stands for commercial and we don't have short forms in Indian languages and hence short forms for TLDs or IDNs will be a real issue and it cannot convey any meaning. Example the word (inaudible) in Tamil means mile while in Marathi it will mean "lizard". So these are some of the issues which will translate if you do transition of TLDs.

Another solution is to translate the TLDs into different languages, however since the TLDs do not convey the language information, it is likely that a translation suitable for one region may not be suitable for other because of the regional translation requirements. This issue is more specific where the scripts or languages are shared across the borders. And these are issues of something called expected display something like the reordering has not happened over there, but the expected display. So there are certain application related issues also. This is one example of an

application issue. This is something rendering again where there is unnecessary spaces which are common in the browser.

Okay, so what we have done is we are actually all these Indian languages – 19 out of 22 – are having a well structured form structured and we are (inaudible) which is called ABNF. This policy is uploaded onto our website and one can download and have a look at the policy for definition of the variants and how we are managing the variants. I will skip this, thank you.

Dennis Jennings: Thank you very much indeed. And again, apologies for having people rush in 10 minutes through hugely complicated subject. Again, we see some of the extraordinary complexity that we're going to try and look at and find what the user expectation is when dealing with domain names, TLD and domain names in IDNs. Our last speaker is Dr. Cary Karp who is Director of the Internet Strategy and Technology at the Swedish Museum of Natural History. And he's a nice simple topic – he's going to talk about the Latin scripts. Alright Cary.

Dr. Cary Karp: So, this is the one that we all know – hit the next slide, one past that. The one that we all know and if it was the one that we all loved we probably wouldn't even be sitting here. This is the script that's used for a larger number of languages on our planet than any other, however very few of them, not even English are served fully well by the basic 26 letter version of it. Next slide please. I'm

going to illustrate this with Swedish – Sweden is an extraordinarily erudite member of the networking community, has been for a long time – it uses a Latin based alphabet; A-Z.

"W" was actually added very recently because the worldwide web needed to be spelled properly. Seriously - prior to that "W" was a variant of "V" and nobody cared. But thanks to this community, the Swedish alphabet is not 28 but now 29 letters long and it's the last three there that you see that are the unique ones. It's important to note that those three are not decorated versions of "A" and "O", they are atomic letters, they do not decompose, that is not a diacritically marked "A" not "A" and "O" with the irises above it.

A number of additional diacritical marks are used an regarded as diacritical marks too as you might expect the term to be used to accommodate proper names of non-Swedish origin and a few other autographic conventions. They're important but they're regarded as somehow a part from the Swedish alphabet itself. Next slide please. And I want to drive a point home, especially considering variants, that there can be decorative use of diacritical marking and there can be contrastive use of it.

Next slide. For example, in English naïveté can be spelled with or without the diacritical marking; the meaning is absolutely clear. Next slide. However in the case of resume and resume, by adding a mark you are changing the meaning of the word. That one single

addition is contrastive, which is a fundamental I would believe to the discussion of variants; whatever we ultimately determine that to mean. Next slide.

Here for example is the name of the North. This is a fully conceivable TLD label. It is the region inhabited by Denmark, Sweden, Norway, Finland, the Pharaoh Islands and Iceland. However, were we to decorate that firs "O" we'd be changing the meaning of the word completely. The second word is "nerd" as we describe it. So aggregating for example norden or blocking because of norden, the second label, is something I don't think that would occur to any other Swedish speaker. These are two absolutely separate strings. And another important concept is that of decomposition. Can you take a marked letter and convert it into two unmarked letters? Next slide.

Here for example, that umlauted "O" as we would describe it, does not in Swedish autography decompose to "OE". However, if you don't have access to a font that includes the 29th letter of the Swedish alphabet, the alternate autography is in fact the undecorated norden. So perhaps we would need to aggregate the two of them nonetheless, however senseless it might otherwise be. Okay. This becomes critical when talking about proper names. Next slide. Goethe – the author – not even in German can be written alternatively with an umlauted "O" and an antiquated form of the Swedish name – gothe, the second one. Actually the closest

bookstore to my apartment in Stockholm is Gothe's Bokhandel, and I don't think Gothe would ever regard himself as comparable to Goethe in any regard whatsoever although he may sell books by him.

The Swedish Government approached this in 2005 with a set of guidelines, which they said are only guidelines however if they're not followed, we will elevate them to binding directive. And that was, it was nice – saying that any database system that includes proper names that is maintained at Swedish public expense need to accommodate the following character repertoire. Any obligatory table – that's the one of them – many of these letters actually do appear in legitimate Swedish context, not foreign.

But if you look at the row with the "I"s in it you'll see an "I" without a dot, you'll see an "I" with a dot, you'll see an "I" with two dots. These are separate and distinct letters. The notion of aggregating because a base character is contained in one string and another is just a patent absurdity. The nice part about this table is that this is in fact all of the Latin letters in the Unicode Code chart that are represented at a single Code Point – you don't need to know what a Code Point is okay.

And there is an optional second table – if we can have the next chart – that these systems might wish to apply. And if you note there we have an "I" with a dot on top of it and an "I" with a dot on

the bottom of it. These are supposed to be entirely foreign to Swedish – this is largely to accommodate Vietnamese. However there are letters here which are writing all Greenlandic so it is a Nordic concern.

The nice part there again is having both of those two tables available in your permissible IDN repertoire obviates any discussion of what one might want to have – it is the only bunch of such letters that are available in Unicode and Unicode defines the Universe. Okay, next one. Now, there are five official minority languages in Sweden. Swedish itself only recently acquired any legal status at all. It's not legally the main language of the country. And of the five minority languages two, Sami and Romani, are not individual languages; there a groups of languages.

So, if we were to want, which is a perfectly reasonable thing to want, a set of names Sweden explicitly, in the names that have legal status in the country we would get the following list – next slide. Okay. I will spare telling you which of these names is – this is Sweden and it's the fourth from the bottom "sverige", that's the Swedish name for Sweden. And then we have other languages. This is a reasonable aggregate, a reasonable variant set – whatever terminology we're going to settle on – for the Swedish Government to put forward as its stake in the – well this is our name and these are our languages. It's not an IDN issue.

There are three things here that are IDN labels, all the rest are flat ASCII. And in fact there's another minority language which doesn't use the Latin script at all, it uses Hebrew script – Yiddish. So the set of names of Sweden is two scripts, some ASCII, some IDN Latin otherwise, and that's a reasonable aggregate. But if we take another look – the Sami languages. Here are three of about – depending on how you count, at least six, maybe twice as many again Sami languages names of Sweden. These are the three that are common in Sweden, there are others common in other Nordic countries – the Sami territory spans the Scandinavian peninsula and goes into the Kola peninsula, which is a part of Russia, and in fact you would expect to see a Cyrillic representation in that aggregate. Next slide please.

And here are four of many Romani languages, some of which are also written in Cyrillic. And they might also wish to have their full set of names. So, if we look at the vertical structure these are the names that Sweden might wish to aggregate. And then we have two horizontal slices there – these are the names that two of the major participating minority communities in the country might wish to have. How do you deal with a situation – that's actually my last slide in fact – where a given label appears with equal legitimacy in several, what would be variant aggregates?

So the whole thing is multiplied in an additional dimension here. And this is Sweden, which doesn't gripe much; we're comfortable

with .se, but if as is now happening, Latin, which was barred arbitrarily from the Fast Track process, now has to be accommodated in the policy basis for the steady state ccTLD IDN, and certainly for the gTLD space, if the Swedish case is as intricate as this how intricate more are the complicated cases going to be. This is an easy one. But again the relief is provided in the fact that the character repertoire is given. That's not the issue, but as it appears in the variant space certainly is.

Dennis Jennings: Well okay, thank you very much. If you'd stay here and I'd ask the other speakers to come up to the front, and Francisco maybe you'll sit down in the front row – and you stay here because I need you to monitor the remote participation. So can our other speakers come up here? And we open the floor to questions from the audience. So while the speakers are coming up, just to remind you what this project is about – it's to try and identify what the user expectation is when we have IDN variants. So questions – please come to the microphone. Please identify yourself and please ask your question of one or all of the speakers.

Chris Dillon: Okay, my name is Chris Dillon and I come from University College London and I would like to ask a question – and I think he's gone of Sheldon Lee about the Chinese case. Now, I noticed under the Fast Track that mainland China was given .[jungor] both in simplified characters and in traditional characters and I think during your talk you said this is a sort of a temporary solution.

That basically if there is a situation where a character has two forms, but actually it may end up being given both of them either I think the SC form was the main one and the TC form was the variant. But I was a bit intrigued because you hinted that that may not be an approach that could be continued with.

Dr. Xiaodong Lee: I think it doesn't mean that is a temporary solution. Currently we use the standard of IETF, we delegate the two different domain names simplified and traditional into the same (inaudible) to ensure that users have the same experience of really the simplified and traditional Chinese. But I mentioned in my presentation it's not perfect, it's too (inaudible) so it will bring some trouble for each user. But even in the future we have a solution; I think it would be transferred from current solution to the future one. So don't worry about that. It's in the near issue.

Chris Dillon: It would have been rather interesting if it had been a situation where there was more than one simplified character because in many cases actually China has one set of simplified characters and Japan has another. In that particular case, the two simplified characters are the same, but there are other situations where they are different. So it would be rather interesting to see what happens if something like that comes up.

Dr. Xiaodong Lee: Yeah it is also my question. But however it comes out with some people from Japan, but you can read the (inaudible). They don't

seem to be have variant issues. But they should be discussed with some Japanese experts deeply. Sure, that so many simplified Chinese characters in Japan and in China...

Chris Dillon: They're usually the same, but most of them...

Dr. Xiaodong Lee: Yeah, many say that, but not every Chinese character is used in Japan now. There's so many characters is very similar, all sim because the Unicode between the CJK unification, that means all of the characters in the CJK unification scope is the same character in China and Korea and Japan.

Chris Dillon: Indeed yes. Thank you very much.

Dennis Jennings: Thank you for the question. We have a speaker at the back, please identify yourself.

Andrew Sullivan: Hi, my name is Andrew Sullivan and I guess today I'm wearing my IETF DNSEXT Working Group hat. I wanted to ask, I guess this is probably a question to the panel, there were a number of examples here that were cases of Unicode decomposition versus composed character issues, some of which are potentially solved just by the IDNA works or possibly by policy issues.

There were other issues that had to do with user interface and the way that different applications do things on different platforms.

And then there were cases that were potential user confusion that was possibly phishing issues or the like. And I would like, maybe, if we could sort out which ones of these things are the things that we think we could actually tackle versus the ones that we just don't have any power over.

In particular I'm wondering if people might comment on the potential for users to adapt to some of these cases. We have the example in English for instance of color and colour; all of those kinds of things. And users have just learned to accommodate themselves to that. Some of these cases I think maybe users are going to be able to do – we saw the example that .in has decided not to worry about those cases. So if you could say a little bit more about which one of these cases you think are truly and deeply impossible versus the ones that you think users are going to be able to cope with, that would be very, very helpful, at least from my point of view.

Dennis Jennings: Thank you for the question. Who'd like to pick that up? Please.

Dr. Sarmad Hussain: Okay, so as you would probably guess it's a gray area and at some point arbitrarily some line has to be drawn. And the question is how far down or up you move that line. I think probably one of the reasons for this, or motivations for this project I guess is to actually start formalizing that and look at script specific perhaps policy other than a generic policy which probably wouldn't be

possible. So yes, that's a good question – where does one draw the line.

I think one of the criteria eventually is not going to be just linguistic. A lot of motivation is going to come from the security and stability criteria I think. And that's really going to be eventually, going to be the deciding factor. So that's going to define the baseline and then one can see how much more up you could move, but you would not be able to go below that threshold which starts infringing upon the security and stability of the system.

Again, these are things that we are all, we have been grappling with from my own experience within the Arabic script community. We have had a group, it's called Arabic Script IDN Working Group, in which we discuss many of these things for many, many days. And I possibly look forward to the opportunity within the community to discuss this and try to come up with a current definition for this. If you are looking for an answer, I'm sorry I can't give you one right now. But I think again, just to close my comment, eventually the criteria is going to be a stability and security criteria. Linguistics is going to be then on top of it.

Dennis Jennings: Dr. Kulkarni.

UF: Dennis, excuse me. We have two comments online when you're ready.

Dennis Jennings: Thank you very much. We'll just take the response to this question and then go to the online. Dr. Kulkarni.

Dr. Mahesh Kulkarni: I believe we were also studying this variant for a very long time and when we talked about .[parad] as a ccTLD to be given and applied. We looked at three or more definitions of the variants in my presentation also. One was homograph basically and another was homophone; that is similar sounding and similar – this color colour.com and so forth. So we felt that going into the homophonic similar sounding would be opening up Pandora's box and then we decided that we should restrict to the homograph, which is similar looking because that's there the end user is going to see some of the things like the famous example of paypal.com 1 and l – and that is where I think we also refer to the security concerns people should not be taken to a phish site and that's where the priority should lie.

And hence we decided, at least in our country, to say that the definition is more or less clear that it is confusingly similar characters. And when you talked about the Unicode normalization part, that we have built into something called the variant table itself. So we have build the variant table.

So as you was also referring to the same blip or the same shape happening to two different code points, we took it as a variant basically and most of the things, what we felt are – thought Indian languages are very complex in nature and 22 different languages and so forth – we felt that there is still a structure within that and if we look at that structure then homograph should be an ideal situation to handle the variant tables.

Dennis Jennings: Thank you. Naela can I take a comment from the remote participation?

Naela Sarras: Yes, first one. Good afternoon, my name is David Cohen, speaking in my own capacity assuming that Mr. Kolesnikov is correct and in Cyrillic there is no variant issues/difficulties – is this and/or other working groups plan on allowing the languages and/pr scripts who face no variant issues to proceed to the next step. (confusion/ASCII similarity or whichever the next step is determined to be)

Dennis Jennings: A very interesting and very good question that I don't know the answer to because I'm not sure that that is a full and complete statement about the Cyrillic script that will be accepted by everybody. If it is, it's most likely that the next step will be to look at the confusability between similar scripts. But we're essentially looking for advice on first of all the variants, and then moving on to confusability. But the point is well made and I don't have an

answer for that particular question. Do you want to take the second comment from…?

Naela Sarras: Sure. The second one comes from Mohamed al-Bashir – he says I think the Arabic presentation contained examples of mixed signal scripts; does the presenter think if IDN guidelines and registry internal policies in handling variants (bundling or blocking) besides non mixing of scripts could limit the types of similarity confusion?

Dennis Jennings: Excellent question. Dr. Hussain?

Dr. Sarmad Hussain: I'm not sure what he or she is referring to as mixing of scripts because what I was presenting was entirely within Arabic script. Is he referring to or she referring to mixing of languages?

Naela Sarras: It's Mohamed – I think so.

Dr. Sarmad Hussain: So, maybe if you're listening, if you could rephrase that question more precisely I can respond to it.

Naela Sarras: Okay, I'll ask.

Dennis Jennings: Cary you wanted to make a comment?

Dr. Cary Karp: Since reference was made to the IDN guidelines – they specifically require or forbid, depending on how you want to look at it, the comingling of scripts in a given label. So if there is a character from a Unicode name script in that label, and that's something that can be indentified algorithmically, then that is the only script that may appear in that label. You can have different languages in the name but one script – and note script, not language – script in a label.

Dennis Jennings: And just to follow that comment – script as defined as a script in the Unicode table.

Dr. Cary Karp: There's something that's crept into this conversation that perpetually puzzles me and that's since when is there a requirement for a domain name label to be a word in any language.

Dennis Jennings: Well I don't have an answer to that question, but indeed, a good question. Edmon, you're next, can you identify yourself?

Edmon Chung: This is Edmon Chung from .asia. Actually I wanted to add to Christopher and Xiaodong's comments on just I think in terms of temporary or current exception situation, I think the Board actually also mentioned in one of the resolutions, Dennis I think you might can correct me if I'm wrong, but did have a resolution specifically saying that the arrangement for the .china and .taiwan situation is somewhat of an exception and this work in part of coming out of

that resolution as well, that we need to look at a longer terms solution.

When Xiaodong mentioned that there is also a possibility of a technical solution that's even more ideal in the future, I just want to add that the current implementation is something that's susceptible, so far in terms of the deployment, I think the user experience has been good and what has been done is something – I think the experience there could be somewhat of an input into the discussions here as well. So I don't, even though it's "temporary" it's not something that people are having problems with at this point. But of course, there are more ideal scenarios.

But adding to that also is that you look at the .taiwan situation, in fact there are actually additional issues. Right now there are two strings delegated that are actually look – I think one or one more that is a preferred string that would, under certain policies, should be also delegated. So I think those are scenarios which this group, I guess these studies would also look into. So in terms of temporary solutions, yes kind of then that's why we want to work more on it, but technically it's something that still works at this point. So there's no immediate need to make a lot of technical changes; at least for Chinese.

And that brings me to my last note is that you're right, we are working with the Chinese, and the Chinese community has worked

on this problem for a long time and perhaps with all the experience and with all the knowledge and the user experience input into this process, Chinese may be one of those that could come out of the gate, if you will, sooner than at the end of some other issues. Thank you.

Dennis Jennings: Thank you for that Edmon. As currently constructed – I'll just make one comment first – as currently constructed the idea is that we'll do the five cases; there were five case reports. And then we'll do a common issues. So we're not envisaging that one case, one script will move more rapidly than another. But that may be something that comes out of the studies. Xiaodong, do you want to comment?

Dr. Xiaodong Lee: I just want a chance to comment – just want to give you some numbers. Right now there are 357 Chinese domain names .china have (inaudible). And how many inquires per day and how many inquiries – about 15% inquiries in traditional Chinese domain name and 85% is inquired simplified domain name. But up to now there is no complaint about the current administration and the administration policies. That's some numbers, but adding to Edmon Chung's comment just for information.

Dennis Jennings: Thank you. Now we've got a question from the back – you've been waiting patiently sir.

Dave Crocker: Dave Crocker; Brandenburg Internetworking. The mechanism that allows using or adding attachments to email is called “mime” and it was created about 20 years ago. It’s original motivation was to support international characters in email. This is a topic that I knew nothing about. I know only slightly more now. There were lengthy presentations by experts of the day and I had several reactions to watching those presentations.

One was that this was an incredibly complicated topic; that assessment has not changed today. The second was that the people who were working on this had been working on it for a long time, were very intelligent, very knowledgeable and well intentioned; and that assessment has not changed. And that I was, as a consequence, very happy that I didn’t have to be the one working on this; that assessment hasn’t changed either.

However, the reason I got up was because I was starting to have, actually I have for a while had a concern that the very extreme complexity of this couple with the very real nature of the problem is causing us to miss the fact that solving this problem – it is not clear that solving this problem will actually do something important.

What I mean by that – it was really excellent to have the Swedish example up there and whether it’s the simplest example of this problem, what’s nice is you can get your brain about that category

of the problem, I can't get my brain around the category for the other scripts that we saw, which is to say, that at its simplest this is a horrendously challenging situation.

As Andrew Sullivan pointed out, we have examples in the real world of humans adapting to this kind of a problem. I am hoping that the research you're doing, the discussions you're doing will look very carefully at the question of the difference between the mathematical problem, which is what we saw examples of today, which is to say absolute difference versus the human problem. In addition, if you solve these, and let's assume for the moment hat you solve them 100%, then the question is how much of the larger set of problems on the internet has been solved.

For example, if you solve all of these character problems will we stop phishing; will we stop abuses on the net? The answer of course is no, but maybe this will solve 50% of that. Well, I doubt that. And yet what is certain is that any mechanisms that deal with this space are going to be complicated. We know that when we create complicated mechanisms we create new problems and they cost a lot.

Consequently, I am hoping that as choices are made, very serious consideration of the cost versus the benefit and the problem versus the mechanism, or the underlying technical detail versus the larger social detail is balanced a lot more carefully. Over in the IETF, I

know, we often suffer from not paying attention to these larger issues. And I'm hoping in the ICANN forum they get better balance.

Dennis Jennings: Thank you for the comment. Cary, I think you want to respond.

Dr. Cary Karp: You moved from commenting on moving the comfortable attributes of the mathematical facet of this, but you moved into a very narrow next door facet as though that were the purpose of all of this. The purpose of this is not to obviate the kind of anguish that you're rightly concerned about. The purpose of this is to recognize the fact that the user community is broadly poly-blocked, and a whopping fraction of that population would probably find any situation illustrated in ASCII infinitely more confusing than you found the scripts here.

So there is a cultural substrata here, that's what we're building on. And it creates problems. We're not simple doing things only if they solve a problem that we know. We are doing this because we have to. Aware of the fact that it is equally likely to generate unforeseen problems that may actually eclipse the kind of things that you're quantifying, but it's something that simply has to be done.

Dennis Jennings: Yes, the next four billion users – when we go from two billion to six billion users, the next four billion will be using personal mobile

devices in languages that I certainly know nothing about and in scripts that I know nothing about. Any other comment on that? Okay, next question please.

Male: (inaudible), former member of APNIC and also APNIC-EC, and also (inaudible). I was told to give a brief introduction to the IFC, how many of you are familiar with IFC 3743, but basically I just give some a design principle why they come out with this IFC. The original idea for TSE basically, logically it was an exception case because the original idea of IFC was try to implement in the protocol layer.

But unfortunately we didn't make it because we was told this stuff should be moved into a registration process; that's why we moved into a registration process and come out with IDN guideline. But based in our capacity because of the IFC 3743 was made by China, Taiwan, Korea, and Japan. So basically we all capacity only focus on Han character.

So the major contribution I can see from the IFC, whether it is useful or not I'm not sure, but major contribution come from the IFC was three points. The first point is validation because initially they introduced a validation concept to the IDN. So we need to set some limit sets for IDN registration. The second contribution is we need to respect our localized normalization – well I cannot say

normalization because basically it doesn't come from Unicode Consortium, but we think that's definitely needed.

So in the IFC we call it preferred call point, and it's this point that's prohibiting is called variant, host set of variant. So, based on this scenario there's three columns of table was created that's come out with the IANA table repository. So last of the scenario, when we come out with the table – was the table useful based on our capacity for the Han character. But I'm not sure whether it's scalable enough to withstand to all kind of language.

Now, based on my understanding when I reviewed it on the table from IANA, eventually I found a lot on the table maybe not so effective. I'm not sure why they tried to implement that kind of table because it doesn't combine with the IFC principle. That's what I'm going to say. Thank you.

Dennis Jennings: Thank you very much for that. I'm going to have to draw this to a conclusion because we're running out of time. I'm sure we could spend a lot more time on this, and indeed I wish we had much more time for each of our presenters. Just to remind you this is the launch of the IDN variant issues project. There's a draft project plan on the web. We're looking forward to your comments and suggestion to make the final initial project plan so we can kickoff this project.

The first thing we're looking to do is to recruit teams of experts – linguistic language and user experts in the various scripts to tell us what the users expect the system to do; what do they expect when they enter a domain label; what do they expect when they enter an email address with a domain end part of it; what do they expect of when there are variants. Because we need to know what the problem is before we can begin to address what the solutions might be.

And one final word, there is an expectation and there has been an expectation that somehow this is a technical problem that the technical people ought to solve. Well, if it is only solvable by for example, changing the DNS protocol, then that is something that doesn't happen on a time scale that would meet anybody's expectations.

The DNS is a very loosely coupled system. So if there were a requirement in the DNS protocol that was required to be everywhere, then we're talking about a long, long, long, long time before we can guarantee that every part of the DNS will respond in the appropriate fashion.

So, do not think that this is something that you can say to the technical people you ought to solve this, because that may not be any solution to the problem. However, let's go back up. We're

looking to start a project; we're looking to recruit lots of people; we're looking for the community to work on this project.

And our project is to support the community, arrive at a definition of what's expected from the system in the area of IDN variants. Can I ask you to thank our speakers who attempted, and attempted very succeeded in 10 short minutes giving us some idea of the complexity. Gentlemen, thank you very much indeed.

[End of Transcript]